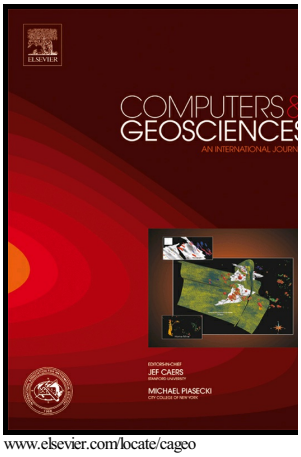
Accelerating Monte Carlo Markov chains with proxy and error models

Accelerating Monte Carlo Markov chains with proxy and error models

Laureline Josset, Vasily Demyanov, Ahmed H. Elsheikh, Ivan Lunati



www.elsevier.com/locate/cageo

PII: S0098-3004(15)30011-X
DOI: http://dx.doi.org/10.1016/j.cageo.2015.07.003
Reference: CAGEO3581

To appear in: *Computers and Geosciences*

Received date: 26 November 2014
Revised date: 17 April 2015
Accepted date: 7 July 2015

Cite this article as: Laureline Josset, Vasily Demyanov, Ahmed H. Elsheikh and Ivan Lunati, Accelerating Monte Carlo Markov chains with proxy and error models, *Computers and Geosciences* http://dx.doi.org/10.1016/j.cageo.2015.07.003

Accelerating Monte Carlo Markov chains with proxy and error models

Laureline Josset[a,*], Vasily Demyanov[b], Ahmed H. Elsheikh[b], Ivan Lunati[a]

[a] *ISTE, University of Lausanne, Switzerland*
[b] *IPE, Heriot-Watt University, Edinburgh (UK)*

Abstract

In groundwater modeling, Monte Carlo Markov Chain (MCMC) simulations are often used to calibrate aquifer parameters and propagate the uncertainty to the quantity of interest (e.g., pollutant concentration). However, this approach requires a large number of flow simulations and incurs high computational cost, which prevents a systematic evaluation of the uncertainty in presence of complex physical processes. To avoid this computational bottleneck, we propose to use an approximate model (proxy) to predict the response of the exact model. Here, we use a proxy that entails a very simplified description of the physics with respect to the detailed physics described by the "exact" model. The error model accounts for the simplification of the physical process; and it is trained on a learning set of realizations, for which both the proxy and exact responses are computed. First, the key features of the set of curves are extracted using functional principal component analysis; then, a regression model is built to characterize the relationship between the curves. The performance of the proposed approach is evaluated on the Imperial College Fault model. We show that the joint use of the proxy and the error model to infer the model parameters in a two-stage MCMC set-up allows longer chains at a comparable computational cost. Unnecessary evaluations of the exact responses are avoided through a preliminary evaluation of the proposal made on the basis of the corrected proxy response. The error model trained on the learning set is crucial to provide a sufficiently accurate prediction of the exact response and guide the chains to the low misfit regions. The proposed methodology can be extended to multiple-chain algorithms or other Bayesian inference methods. Moreover, FPCA is not limited to the specific presented application and offers a general framework to build error models.

Keywords: functional data analysis, Bayesian inference, two-stage MCMC, Imperial College Fault test case

*Corresponding author. ISTE, University of Lausanne, Geopolis - UNIL Mouline, 1015 Lausanne, Switzerland. Tel.: +41 21 692 44 18.

Email addresses: laureline.josset@unil.ch (Laureline Josset), vasily.demyanov@pet.hw.ac.edu.uk.co (Vasily Demyanov), ahmed.elsheikh@pet.hw.ac.edu.uk.co (Ahmed H. Elsheikh), ivan.lunati@unil.ch (Ivan Lunati)

1. Introduction

Simulations of subsurface flow is important in many applications, such as groundwater protection and remediation, water prospection, exploration of hydrocarbon resources, and nuclear waste disposal. One of the main challenges is to estimate a continuous distribution of the underground model parameters from a sparse set of observational sites. This lack of information on model input propagates to the quantities of interest (for instance, the concentration of a pollutant in a drinking well), whose exact values remain uncertain. Model calibration using historical integrated data (for example, time series of concentration or pressure at observation wells) is often used to reduce the uncertainty on model parameters by relying on Bayes theorem. A widespread approach for numerical application of Bayes rule is to use Monte-Carlo Markov-Chain (MCMC) simulations (Robert and Casella, 2004) to sample the posterior probability density function. While MCMC is theoretically robust and ensures convergence to the true posterior distribution under mild constraints, in practice it is subject to several limitations due to the cost of the large number of required flow simulations, which can become prohibited in presence of limited computational resources. Indeed, the finite length chains should be able to explore all areas of the prior space in order to provide samples from the posterior distribution. To achieve this goal, it is tempting to increase the step length of the chains, but this would result in a drastic reduction of the acceptance rate (which should ideally remain around 20-50% in multidimensional space) and subsequently in a high number of wasted simulations (Roberts et al., 1997).

To avoid these issues, Efendiev et al. (2005, 2006) and Christen and Fox (2005) have introduced a two-stage MCMC, which employs a less computationally expensive solver to obtain a first evaluation of the proposal and decide whether it is useful to run the exact solver. This allows them to reduce the number of exact simulations that will be rejected and thus increase the acceptance rate. This methodology has been first explored by Christen and Fox (2005) to recover resistor values of an electrical network from measurements performed at the network boundary. They have obtained an increase in acceptance rate (the number of exact simulations accepted over the number of exact simulations run; first-stage simulations are not taken into account as their cost is assumed to be negligible). Both Efendiev et al. (2006) and Christen and Fox (2005) have shown that, under certain hypotheses, the solution converges to the posterior distribution. Efendiev et al. (2005, 2006); Dostert et al. (2008) have applied this methodology in the context of flow in porous media. As first-stage solver they have used a multiscale method, which combines a global coarse solution with a number of local fine solutions. If the coarse solution is accepted, local solutions are employed to reconstruct a finer solution on the original grid, based on which the second-stage evaluation is performed. While this allows for the necessary convergence assumptions to be satisfied (namely, smoothness and strong correlation), the computational gain of the two-stage set-up is limited. Indeed, the reconstruction step

(necessary for the second-stage evaluation) is cheap with respect to the cost of constructing and solving the coarse problem used at the first-stage. Other applications of two-stage MCMC have used polynomial chaos response surfaces (Zeng, 2012; Elsheikh et al., 2014; Laloy, 2013) as first-stage model. The computational gain is much higher, despite some additional cost required to set up the polynomial chaos model.

The use of inexact solvers requires designing error models to account for the discrepancy between approximate and exact responses. In the context of multiscale approaches, Kennedy and O'Hagan (2001) used a Gaussian-process method to represent model inadequacy. O'Sullivan and Christie (2005, 2006) employed error modeling to reduce the bias in history matching resulting from the use of upscaled reservoir models. Efendiev et al. (2009) proposed non-linear error models in the context of ensemble-level upscaling. Scheidt et al. (2010), for instance, used a distance metric to account for upscaling errors in ensemble history matching. More specifically to two-stage MCMC, Cui et al. (2011) proposed to adapt the error model at each iteration: they used information on the discrepancy between the exact and approximate models at the previous iteration to correct the result of the successive iteration. However, this approach works and provides a good correction only for problems that are smooth enough.

Here, we propose a different strategy that combines a two-stage MCMC set-up with a methodology recently presented by Josset et al. (2015). We use an approximate model (proxy) that assumes a very simplified physics with respect to the problem under consideration, and we construct an error model to account for the approximation errors. The error model is purpose oriented as it is tailored directly for the quantities of interest following an approach typical of machine learning. For a subset of realizations, the responses of both the proxy and the exact models are evaluated and the mapping between the two is learned by means of tools from functional data analysis (Ramsay, 2006; Ramsay et al., 2009). Josset et al. (2015) applied this methodology to propagate the uncertainty on the permeability field to the concentration of a pollutant in the observational well. Here, the methodology is tested on a complex problem of Bayesian inference, the Imperial College Fault (ICF) test case, which is a benchmark problem first published by Tavassoli et al. (2004) and repeatedly explored in many studies (e.g., Demyanov et al. 2010; Mohamed et al. 2011, 2012).

The paper is structured as follows: we first describe the ICF test case and review the literature about the calibration of this model (Section 2). Next, we present the novel methodology, which uses a purpose-oriented error model within a two-stage MCMC set-up (Section 3). Then, we specifically construct and evaluate the error-model approach for the ICF problem (Section 4.1). Finally, we compare and discuss the results of the two-stage MCMC with the classic Metropolis-Hastings algorithm (Section 4.2).

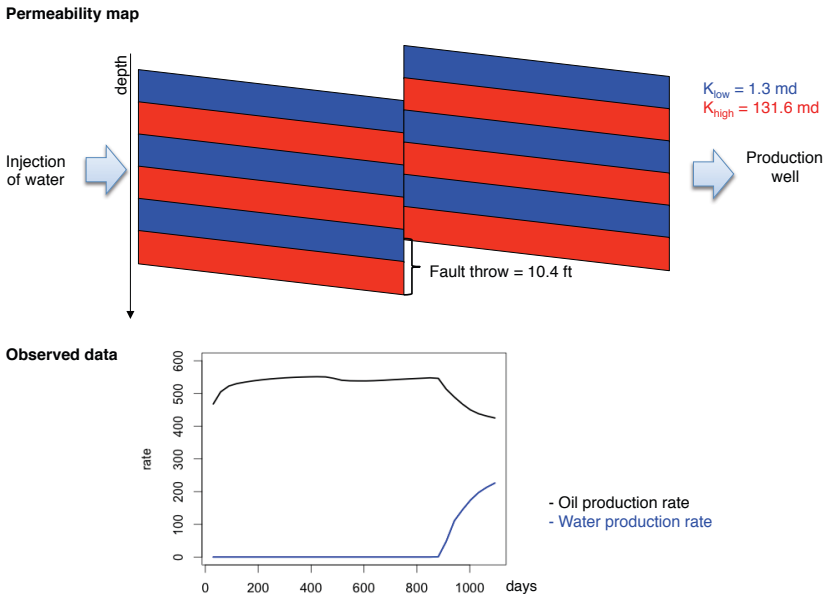


Figure 1: The permeability map of the ICF test case and the observed data used for the history matching. As prior, a uniform distribution is attributed to each parameter, i.e., $P(h) = \mathcal{U}_{[0,60]}$ for the fault throw h, $P(K_{high}) = \mathcal{U}_{[100,200]}$ for the permeability of the most permeable facies K_{high}, and $P(K_{low}) = \mathcal{U}_{[0,50]}$ for the permeability of the least permeable facies K_{low}.

2. The Imperial College Fault (ICF) test case

The ICF test case was first published by Tavassoli et al. (2004, 2005) as a simple yet challenging example of history matching in petroleum engineering applications. Since then, ICF has proved a difficult test for optimization techniques due to numerous local minima. The ICF model consists of a layered reservoir disrupted by a fault (figure 1), in which water is injected at the left-hand boundary while the displaced fluids are recovered at the right-hand boundary. The layer-cake model of the reservoir permeability is described by three parameters: the conductivity of the high permeability facies, K_{high}, the conductivity of the low permeability facies, K_{low}, and the fault throw, h. The true parameters are $K_{high} = 131.6$ md, $K_{low} = 1.3$ md and $h = 10.4$ ft. A uniform distribution $\mathcal{U}_{[a,b]}$ (where a and b are the bounds of the distribution) is attributed to each parameter as prior.

The calibration of the parameters to the observational data (oil and water production rates) appeared to be a challenging history matching problem. Due to the nature of the permeability field, several parameter combinations, corresponding to narrow regions of the parameter space, can reproduce the observational data with satisfactory accuracy. Between these regions of good quality, the misfit is very high due to the very irregular response surface that results from the strong fluctuations of the connectivity across the fault when h is varied. We refer to figure 9 for a 1D cross-section cut of the complex misfit surface that characterizes

this problem.

Many optimizations and inference techniques have been applied to the ICF problem over the years. The first studies of this test case (Tavassoli et al., 2004, 2005; Carter et al., 2006) have employed a pure Monte Carlo approach, which required nearly 160'000 samples of the parameter space. Christie et al. (2006) demonstrated that a good representation of the uncertainty can be inferred from a few thousand samples using Genetic Algorithm Important Sampling with artificial neural network proxy. More recently, Demyanov et al. (2010) have used Support Vector Machines (SVM) with a small number of flow simulations (about 700); and Mohamed et al. (2011) have employed Particle Swarm Optimization (PSO) using 2050 flow simulations. A Bayesian inference approach close to two-stage MCMC has been presented by Mohamed et al. (2012), who used a population MCMC method with 45'000 simulations. We refer to Mohamed et al. (2011) for a more detailed review of the literature on the ICF problem.

3. Methodology

Our objective is to sample the geostatistical parameter space conditioned on some flow observations. Using Bayes theorem, this can be written as

$$P(k|d) \propto P(d|k)P(k) \tag{1}$$

where $P(k|d)$ the is probability of the realization with the parameters, k, conditioned on the data, d, and $P(d|k)$ the likelihood distribution. The most common technique to tackle this problem uses the Metropolis-Hasting (MH) algorithm (Robert and Casella, 2004), which is very demanding in terms of CPU time. We propose to employ a two-stage MCMC algorithm in which the first stage allows us to reject samples from low likelihood regions of the parameter space based only on the responses of an approximate model. The latter is constructed by combining a proxy model with an error model that permits the reduction of the proxy bias. This approach is illustrated in figure 2.

3.1. Error modeling based on Functional Principal Component Analysis (FPCA)

The number of flow simulations required for MCMC or two-stage MCMC can become prohibitive in case of very complex physical processes that requires performing computationally expensive simulations. An inexpensive proxy that relies on a very simplified physical description can be used to reduce the computation cost. However, direct inference from the proxy response is extremely dangerous, because the proxy model neglects important physical couplings inherent to the system, which likely bias the predictions. However, if we are able to devise an effective model of the errors arising from the use of the proxy, we can account for the neglected complexity and correct the bias of the prediction.

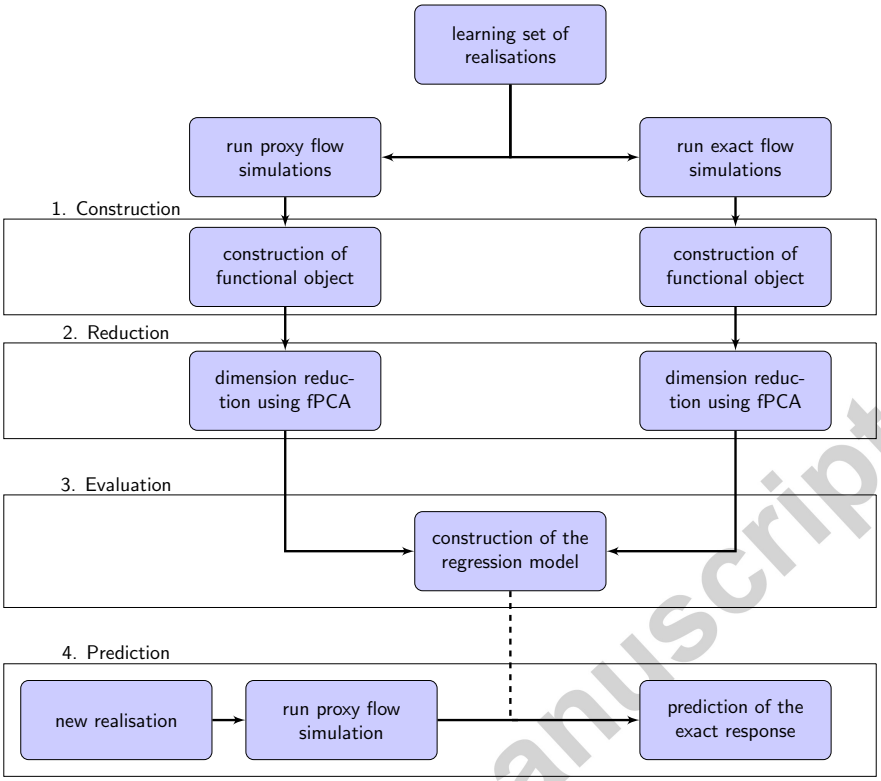


Figure 2: Flowchart of the construction of the error model as proposed in Josset et al. (2015). Numbering refers to the sub-sections in section 3

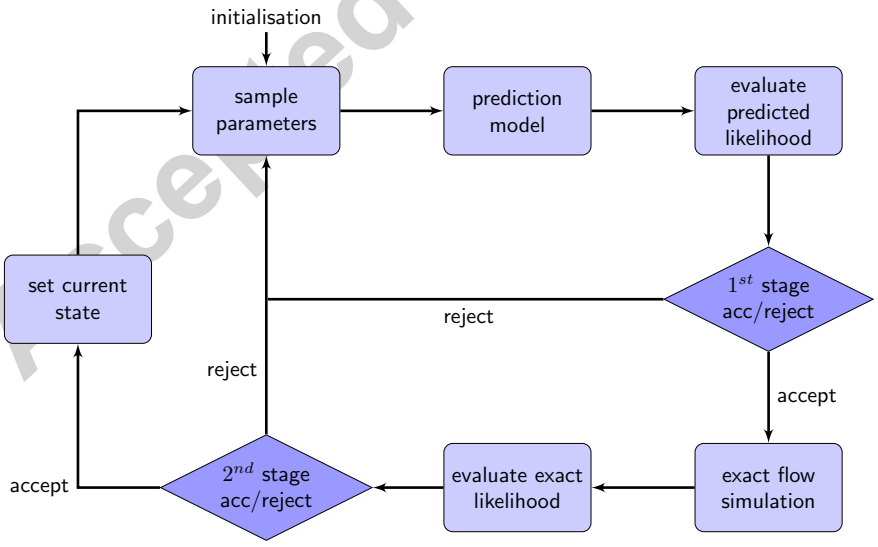


Figure 3: Flowchart of the two-stage MCMC algorithm.

A purpose-oriented error model can be constructed directly on the quantity of interests by training a regression model on a subset of response pairs obtained by evaluating the proxy and the exact model for a selected subset of realizations (Josset et al., 2015). The flowchart of the regression-model construction is detailed hereafter and illustrated in figure 2.

3.1.1. Construction of the learning set of curves

The first step consists in constructing the learning set from pairs of proxy and exact response curves corresponding to the same realizations. To obtain a learning sample of N realizations, which is assumed representative of most plausible solutions, we use the Latin Hypercube Sampling (Carnell, 2009). Other sampling methods (e.g., basic random sampling of the prior or stratified sampling) could be successfully employed as long as the various regions of the prior are sampled.

Once the learning realizations are identified, the proxy and the exact solutions are computed to get the time-dependent response curves. The functional proxy curves, $\{x_i(t)\}_{i=1,\dots,N}$, and functional exact curves, $\{y_i(t)\}_{i=1,\dots,N}$, are obtained by interpolating the responses produced by the numerical models, which are discrete in time, by means of a basis of spline functions.

Notice that a functional representation of the curves is necessary to deal with data acquired with different time resolution, as it is always the case when the numerical solvers employ adaptive time stepping techniques. The drawback is that a functional full-regression model between continuous curves is difficult to implement and requires introducing and fine-tuning additional parameters. To avoid these problems we proceed to a functional reduction of the problem dimensionality.

3.1.2. Functional reduction of the dimensionality

We reduce the dimension of the response spaces by means of Functional Principal Component Analysis (FPCA, Henderson 2006), which is a rather straightforward functional extension of standard PCA. Beside the indubitable computational advantages, low-dimensional spaces allow us to visualize the most relevant modes that describe data variability and help us to evaluate the suitability of the proxy model for the quantities of interest. FPCA is applied separately to the two sets of exact and proxy responses. The dimensionality of the response spaces is reduced considering only the first D harmonics, where D is chosen to achieve the desired degree of accuracy.

Although FPCA offers an optimal dimensionality reduction with respect to the total mean squared error, any rotation of the basis preserves the accuracy. The choice of a proper rotation of the basis might allow a better interpretation of the data (Richman, 1986; Ramsay et al., 2009). Therefore, we use the *varimax*

algorithm (Kaiser, 1958) to find an appropriate rotation. As a results, each proxy response is approximated by projection on the rotated FPCA basis as

$$x_i(t) \approx \tilde{x}_i(t) = \bar{x}(t) + \sum_j^D b_{ij}\zeta_j(t), \tag{2}$$

where $\bar{x}(t)$ is the mean curve, and

$$b_{ij} = \int [\bar{x}(t) - x_i(t)]\zeta_j(t)dt \tag{3}$$

is the projection of the deviation from the mean of the i^{th} proxy curve on the j^{th} rotated harmonic $\zeta_j(t)$. Following the same procedure, the N exact responses in the learning set are approximated as

$$y_i(t) \approx \tilde{y}_i(t) = \bar{y}(t) + \sum_j^D c_{ij}\eta_j(t), \tag{4}$$

where $\bar{y}(t)$ is the mean exact response, $\eta_j(t)$ the j^{th} harmonic of the (varimax) rotated orthonormal basis $\{\eta_j(t)\}_{j=1,\dots,D}$, and

$$c_{ij} = \int [y_i(t) - \bar{y}(t)]\eta_j(t)dt \tag{5}$$

the score with respect to $\eta_j(t)$.

3.1.3. Regression and error model

The relationships between the two sets of curves in the learning set approximated is investigated by considering the first D harmonics, $\{\tilde{x}_i(t), \tilde{y}_i(t)\}_{i=1,\dots,N}$. As sketched in figure 4, the goal is to find a mapping from the space of proxy responses onto the space of exact responses that allows us to predict the exact responses for the realizations that do not belong to the learning set (hence, without actually solving the exact model). This is commonly referred to the model's predictive ability.

Here, we restrict ourselves to functional linear regression models that minimize the l_2-norm of the residuals

$$\varepsilon_i = y_i - \hat{T}(x_i) \qquad i \in [1, \dots, N], \tag{6}$$

where $\hat{T}$ is the estimator on the learning set. Training such a functional linear model in full generality is not straightforward, but we can take advantage of the FPCA basis to define a multivariate multiple regression problem of the form (Hastie et al., 2009; Fox and Weisberg, 2010; Weisberg, 2014)

$$c_{ij} = \beta_{0j} + \sum_{l=1}^{D} b_{il}\beta_{lj} + e_{ij} \quad (i,j) \in [1,N] \times [1,D], \tag{7}$$

where β_{lj} are the coefficients of the regression, and e_{ij} are the errors, which we assume to be Gaussian with variance σ_j^2.

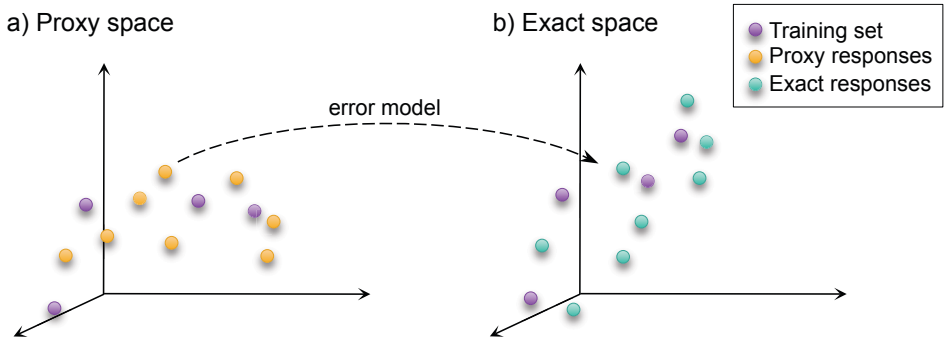


Figure 4: A statistical model is built on the learning set to relate the coefficients of the elements $x_i(t)$ in the proxy space to the coefficients of the elements $y_i(t)$ in the exact-model space. It is used as an error model to predict the exact response from the proxy response.

A further simplification is obtained by splitting the regression model into D independent problems of the form

$$c_i^{(j)} = \beta_0^{(j)} + \sum_{l=1}^{D} b_{il}\beta_l^{(j)} + e_i^{(j)}. \tag{8}$$

This simplification does not affect the operator estimators, which are identical for the problems in Eqs. 7 and 8, i.e., $\hat{\beta}_{jl} = \hat{\beta}_l^{(j)}$. However, confidence bands of the multivariate regression model cannot be directly derived from those obtained for the regressions in equation 8, which complicates their derivation (Josset et al., 2015).

3.1.4. Prediction of exact response from the proxy response

The regression model can be used to predict the exact response of any new realization r for which the proxy response $\tilde{x}_r(t)$ is known. Indeed, the estimator of the linear regression model allows us to predict the scores of exact response curve, $\hat{c}_{rj}$, without solving the exact model. Therefore, solely on the basis of the scores of the proxy responses, b_{rl}, we can estimate the exact response as

$$\hat{y}_r(t) = \bar{y}(t) + \sum_{j=1}^{D} \hat{c}_{rj}\eta_j(t), \tag{9}$$

where

$$\hat{c}_{rj} = \hat{\beta}_{0j} + \sum_{\ell=1}^{D} \hat{\beta}_{j\ell} b_{r\ell}, \tag{10}$$

are the estimates of the exact scores predicted by the error model.

3.2. Two-stage MCMC

Two-stage MCMC has been introduced by (Christen and Fox, 2005; Efendiev et al., 2005, 2006) to improve the acceptance rate of the Metropolis-Hastings algorithm (MH). For optimal convergence conditions

of standard MCMC algorithms it is necessary to tune the random-walk step of the chain in order to obtain an acceptance rate between 20% and 50%. As flow simulations are performed at each step to compute the likelihood $\mathcal{L}$ of the proposed sample ϕ, the low acceptance rate implies that 50% to 80% of the flow simulations are performed on rejected samples and do not contribute to the posterior distribution.

Moreover, in order to satisfactorily explore the prior space under the constraint of limited computer resources, the length of the random-walk step is often increased with the result that the acceptance rate is drastically reduced (for instance, an acceptance rate around 10^{-5} is reported by Efendiev et al. (2005)).

The goal of two-stage MCMC is to decrease the computational cost by reducing the number of full-physics flow simulations that are performed on rejected samples. This is achieved by employing an approximate model to identify samples in low likelihood regions that might be rejected and avoid running the exact simulator on these samples and at the same time to identify the samples that are more likely to be accepted by the exact model. Proposing samples that are more likely to be accepted at the second stage will eventually boost the acceptance rate. In other words, the approximate likelihood $\tilde{\mathcal{L}}$ of the proposed sample ϕ is estimated by using the approximate model response, $\hat{y}_\phi(t)$, from which the first-stage acceptance,

$$\tilde{\alpha} = min\left\{1, \frac{\tilde{\mathcal{L}}(\hat{y}_\phi(t))}{\tilde{\mathcal{L}}(\hat{y}_\theta(t))}\right\}, \tag{11}$$

is computed. If the sample is accepted, the response of the exact model, $y_\phi(t)$, is calculated to compute the exact likelihood $\mathcal{L}(y_\phi(t))$ and the proposal is tested again using a modified acceptance/rejection condition

$$\tilde{\alpha} = min\left\{1, \frac{\mathcal{L}(y_\phi(t))}{\mathcal{L}(y_\theta(t))}\frac{\tilde{\mathcal{L}}(\hat{y}_\theta(t))}{\tilde{\mathcal{L}}(\hat{y}_\phi(t))}\right\}. \tag{12}$$

A schematic diagram of the two-stage MCMC algorithm is depicted in Figure 3.

Efendiev et al. (2006) demonstrated that the two-stage MCMC converges to the true posterior distribution under two mild assumptions: first, the proposal distribution has to satisfy $q(\phi, \psi) > 0$ for any (ϕ, ψ) in the posterior distribution; second, the support of the exact posterior distribution belongs to the support of the approximate distribution (see theorem 3.2 in (Efendiev et al., 2006)).

The first condition is easily satisfied when a Gaussian random walk is used as proposal distribution: a step size sampled from a normal distribution guaranties that $q(\phi, \psi) > 0$ for any (ϕ, ψ). The second condition is met assuming a Gaussian error model for the likelihoods for both proxy, $\hat{y}_\phi$, and exact, y_ϕ, solutions, i.e.,

$$\tilde{\mathcal{L}} \propto \exp(-\frac{||y_{obs} - \hat{y}_\phi||^2}{\sigma_{app}^2}) \qquad \text{and} \qquad \mathcal{L} \propto \exp(-\frac{||y_{obs} - y_\phi||^2}{\sigma_{ex}^2}), \tag{13}$$

respectively. The likelihoods distributions are non-compact, and thus the supports of both posterior distributions are identical to the one of the prior distribution.

Numerically, it is probable that the likelihood values are very close to zero, which prevents the chain to reach all regions of the parameter space. However, under the condition that the exact and approximate

misfits are correlated, Efendiev et al. (2006) have shown that it is possible to choose σ_{app} such that the second assumption is verified and that the optimal acceptance rate can be obtained by setting σ^2_{app} to σ^2_{ex}/α_o, if the correlation can be described by a linear relationship

$$||y_{obs} - y_\phi||^2 \approx \alpha_0 \cdot ||y_{obs} - \hat{y}_\phi||^2 + \alpha_1. \tag{14}$$

4. Application to the IC Fault test case

In this section, we first assess the performance of the functional error model to satisfactorily describe the misfit between the proxy and the exact models for the ICF test case. Then, the proxy (corrected by the error model) is used as first-stage solver in two-stage MCMC, and the results are compared with a pure Metropolis-Hastings approach in order to illustrate the potential of error modeling in the context of Bayesian inference.

4.1. Error model

The objective of functional error modeling is to correct the proxy response to estimate an unbiased exact response. The first step is to choose an appropriate proxy that is sufficiently informative of the behavior of the exact model but considerably cheaper in terms of computational cost.

4.1.1. Choice of proxy model

Here, we are interested in sampling the space of the parameters that describe the permeability field, while the properties of the fluids and the physical processes are known. We consider the simultaneous flow of two immiscible liquids that form two separate phases (oil and water) and we are interested in the production rates of both fluids. Under these conditions, the fluid transport is governed by a set of coupled nonlinear equations, which complicates the numerical solution of the equations. The high degree of coupling between the pressure and the saturation equations renders the transport problem computationally expensive.

A natural choice of proxy is to neglect the nonlinearity of the permeabilities and the two-way coupling between the equations by solving a simple tracer transport problem. This means using a single phase solver as a proxy for a two-phase solver. Further simplifications are introduced by neglecting capillarity and gravity, so that the pressure equation has to be solved only once per proxy simulation.

4.1.2. Construction of the learning set

The construction of the learning set requires making choices on the method of selection and on the size of the set. Here, we train the error model on a subset of 100 realizations selected by performing a Latin hypercube sampling in the 3D parameter space. The learning set consists of two pairs of curves par realization: water and oil production rates obtained with the proxy and the exact models. Comparison with other sampling techniques and learning-set sizes has indicated that the effects of these variables on the error

model is limited. Additional tests (not reported here) have suggested that 20 realizations might be sufficient to obtain a satisfactory error model, but with such few realizations the performances would vary greatly from one learning set to another. The choice of a subset of 100 realizations has been made for the sake of robustness. The proxy and exact curves in the learning set are plotted in figure 5.a.

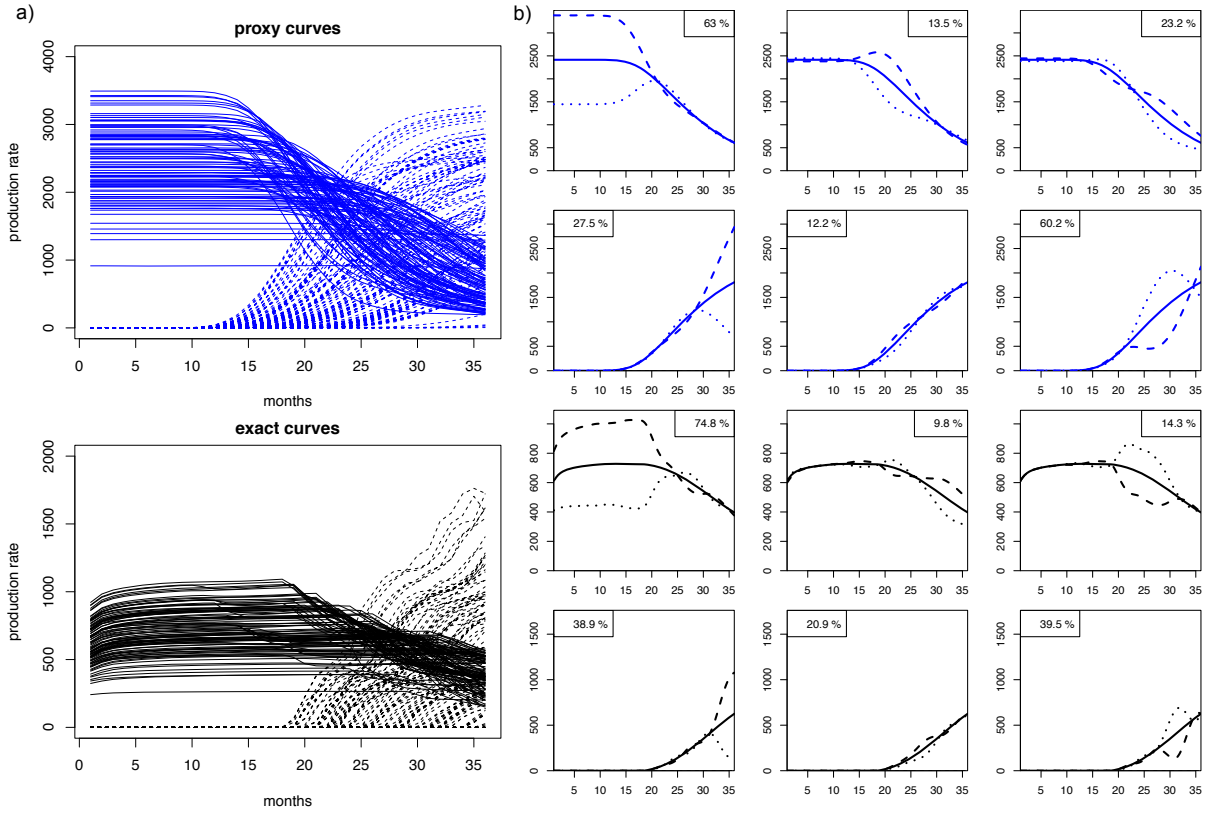


Figure 5: a) The learning set of curves is constructed by running both proxy (top) and exact (bottom) models on the sampled geostatistical realizations. The production rates of oil (full lines) and water (dashed lines) are plotted in bbl/day in function of time. b) The three first rotated functional principal components (harmonics) extracted from the learning set are represented here for the two sets of pairs of production rate curves. The solid lines are the mean curves and the dotted lines represent the variability around the mean described by the corresponding harmonic. The legends report the percentage of the total variability, which is explained by each harmonic.

4.1.3. Dimensionality reduction and interpretation of the information

For each realization in the learning set we have four subspaces of response curves: the spaces of the proxy and exact production rates of water and oil. For each subspace, we subtract the average response from each response curve and then apply FPCA to obtain a basis of the subspace. To reduce the dimensionality of the problem we truncate the basis by considering only the first three functional principal components, which capture more than 96% of the variability within the learning set.

By close inspection of the rotated harmonics (figure 5.b), we notice that the first principal component captures the variability of the initial plateau of oil production rate (i.e. prior to the water breakthrough, figure 5.a bottom). The second harmonic of the proxy and the third harmonic of the exact model describe the production drop after water breakthrough. The third harmonic of the proxy and the second harmonic of the exact model capture the remaining late-time variability. A similar analysis can be done for the harmonics of the water production rate curves. The first harmonics (both of the exact and proxy models) explain the variability at the end of the simulation time, the second harmonics capture small variabilities at the water breakthrough time, and the third harmonics describe most of the variabilities occurring at intermediate time between the water breakthrough and the end of the simulation.

4.1.4. Evaluation of the informativeness of the proxy and self-consistency of the error model

After the dimensionality reduction, each functional space has a six-dimensional basis (three harmonics for the water production and three harmonics for the oil production). In addition to decreasing the computational cost of constructing the error model, the reduction to six dimensions facilitates a visual inspection of the relationships between proxy and exact curves, providing insight into whether the proxy response is informative of the full-physics response.

Figure 6.a) plots the one-to-one relationship between the scores (i.e., the projections on the harmonics) in the proxy space versus the scores in the exact space. A clear linear relationship can be observed in the upper-left plot, which illustrates the relationship between the first harmonics of the oil production. This indicates that the height of the plateau of the exact oil-production curves is well explained by the proxy plateau. On the other hand, the second harmonic of the proxy oil curves (plots in the second column) does not display a simple relationship with any harmonic of the exact curves. Also, the second and third harmonics of the exact oil-production curves do not display a simple relationship with any of the proxy harmonics (second and third rows). This indicates that the proxy is not very informative of the features described by the second and third harmonics of the exact oil curves and one can expect that the error model will be less accurate in predicting those harmonics.

The error model maps the space of the proxy responses onto the space of the exact responses and it is constructed by solving six independent linear regression models as explained in section 3.1.3. Figure 6.b) shows the correlation between exact scores and the scores predicted by the error model (in the space of the exact curves) for all the 100 realizations of the learning set. As expected, the projection on the first oil-production harmonic, which describes the plateau at early time, is well predicted with an R^2 value of 0.91. The projections on the second and third harmonics are predicted with lower accuracy ($R^2 = 0.77$ and 0.79, respectively). The water-production scores are rather well predicted with R^2 values around 0.9. The underestimation of the largest score values for the first and the second harmonics of the water production rates (figure 6.b) demonstrates the limitation of the linear model. Indeed, as the proxy curves are always

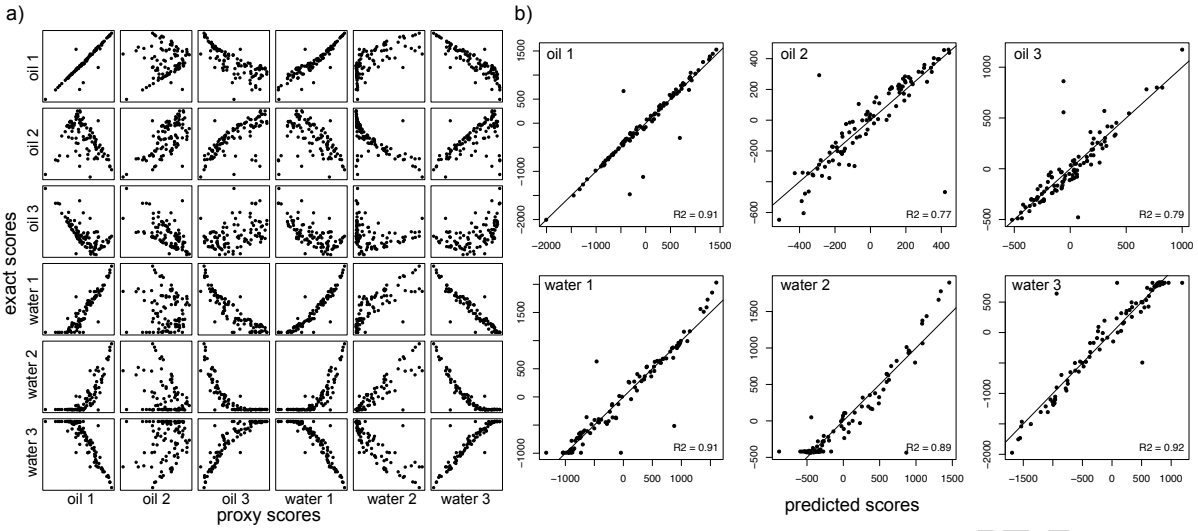


Figure 6: a) Dependency between exact and proxy scores. The scores of the first three harmonics of the exact oil production rate $\{\eta_i^o(t)\}_{i=1,2,3}$ and water production rate $\{\eta_i^w(t)\}_{i=1,2,3}$, are plotted as function of the scores of the proxy curves with respect to the harmonics $\{\zeta_i^o(t)\}_{i=1,2,3}$ and $\{\zeta_i^w(t)\}_{i=1,2,3}$. b) Results of the linear model: the exact scores are plotted as function of the predicted scores; also shown is the identity line. Both plots are helpful to assess whether the linear regression model is appropriate to describe the relationship between proxy and exact scores, thus the level of informativeness of the learning set.

positive, not all scores values are possible. In particular, for the second water harmonic (figure 6.a), a clear lower bound in the exact scores is displayed and biases the linear regression.

4.1.5. Evaluation of predictive power of the error model

For a new point in the parameter space, the corresponding realization is built and the proxy model is run. Then, from the output of the proxy model (i.e., the time-discrete recovery rates resulting from the numerical simulations), continuous oil and water production rates are reconstructed and projected on the harmonics. The proxy scores are used as input of the error model, which allows prediction of the corresponding exact scores that are used to reconstruct the two-phase response curves.

In order to evaluate the performance of the error model, proxy and exact simulations were run for a test set of 1000 realizations sampled in the entire parameter space by means of Latin Hyper Cube sampling. Figure 7 compares the exact responses with the predicted responses for four points sampled in the parameter space. Figure 8.a) plots the error of the prediction as a function of time. The error of the mean of the predicted curves is very close to zero for both the oil and the water production rates, which indicates that the predicted mean is not biased. The histograms in figure 8.b) show the distribution of the l_2 and l_∞ error norms. On average, the maximum error made is around 80bbl/day for oil and 180bbl/day for water, respectively.

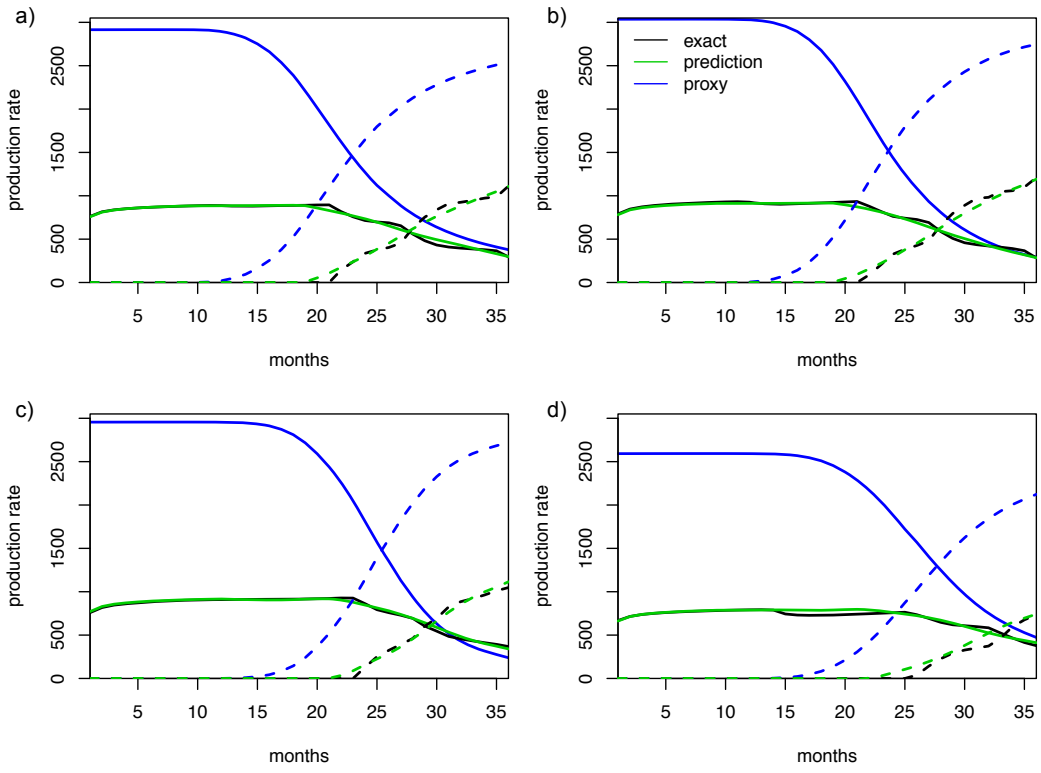


Figure 7: Four predictions that are representative in term of l_2 error norms: a) and b) have errors close to the median, c) to the 25% percentile, and d) to the 75% percentile. The continuous lines are the oil production rates, the dashed lines the water production rate. The proxy curves (blue) are effectively corrected by the error model and the predicted curves (green) match well the exact curves (black).

In the context of Bayesian inference, a correct prediction of the misfit to the observed data is crucial. Figure 8.c) illustrates the correlation between the misfit computed from the predicted curves and the misfit computed from the exact curves for the observational data shown in figure 1. The overall correlation between the exact and predicted misfits is good as indicated by the high correlation coefficients in R^2. Therefore, the prediction model is expected to be efficient at rejecting realizations. However, for small misfits (i.e., for realizations whose responses deviate less from data) the error model is less accurate and tends to overestimate the misfit. This explains the lower Kendall correlation coefficient (a measure of rank correlation) with respect to the Pearson coefficient (a measure of the degree of linear dependence).

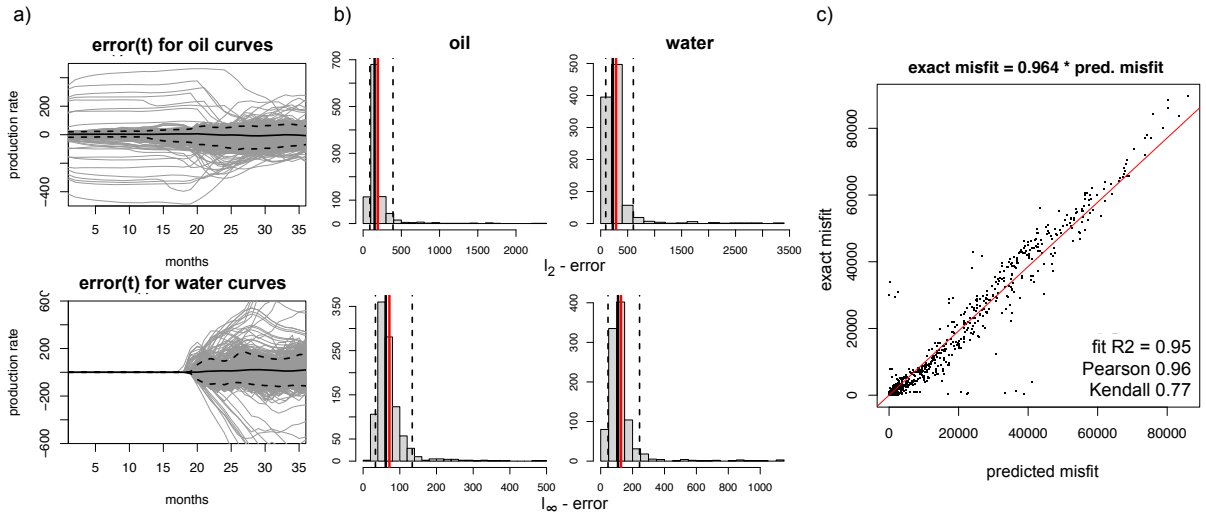


Figure 8: The quality of the error model is evaluated on a test set of 1000 new realizations. a) Difference between production rates predicted with the error model and the exact production rates (grey curves) for the oil (top) and the water (bottom). b) Histograms of the l_2 and l_∞ error. c) Exact misfit versus predicted misfit with respect to the observations (the identity line is plotted in red); the R^2, Pearson and Kendall correlation coefficients are reported to indicate the quality of the prediction.

4.2. Two-stage MCMC

In this section, we first introduce the definition of the misfit necessary to compute the likelihoods in Eq. 13; then we investigate the fidelity of the response surface predicted by the error model; and finally we show that a two-stage MCMC set-up is able to explore a larger portion of the parameter space than MH at the same computational cost, which can be a substantial advantage for challenging problems as the ICF test case.

4.2.1. Definition of the misfit and response surfaces

Here we employ the definition of the misfit that is commonly used to investigate the ICF test case, i.e.,

$$\mathcal{M}_j = \sum_{i=1}^{36} \frac{(C_o^j(t_i) - C_o^{ref}(t_i))^2}{\sigma_o^2(i)} + \sum_{i=27}^{36} \frac{(C_w^j(t) - C_w^{ref}(t))^2}{\sigma_w^2(i)} \tag{15}$$

where $\sigma_o(i) = 0.03 \cdot C_o^{ref}(t_i)$ and $\sigma_w(i) = 0.03 \cdot C_w^{ref}(t_i)$. The likelihood is then obtained from the misfit as $\mathcal{L} = \exp(\mathcal{M}_j)$. Notice that only the water production rate at later time ($i \geq 27$) contributes to the misfit.

The three first original papers on ICF (Tavassoli et al., 2004, 2005; Carter et al., 2006) have employed a slightly different definition of the misfit, which considers the contribution of the water production rate at any time (i.e., with $i = 1$ instead of $i = 27$ in the second summation in Eq. 15). However, this choice leads to a very discontinuous response surface, for which hardly any method beside classical Monte Carlo would be able to provide a reasonable solution. The modified misfit function defined in equation 15 has been introduced to make the problem more tractable and is commonly used in all investigations of the ICF test case.

4.2.2. Comparison of the response surfaces

To further assess the performance of the error model, figure 9 compares the 1D response surface of the misfit of both the exact model and the prediction given by the error model, as a function of the fault-throw value. The response surface of the exact model exhibits several local minima separated by large misfit regions. This situation is particularly challenging for any MCMC approach because many realizations are required to cross large misfit regions with small random-walk steps.

The predicted response surface (which provides the basis of the first-stage rejection decision) is in excellent agreement with the exact response surface for $h > 48$ft. For a fault throw between 8 and 48ft, the discrepancies between the two curves are more important, but the main features of the curves are reproduced. We can expect that the low misfit values of the predicted response curve will be able to guide the chain into this region. For values between 0 and 8 feet, the misfit is greatly overestimated but the shape of the curve is reproduced. If inference is made only based on the prediction model, the minimum around 7ft would not be identified. However, in a two-stage set-up the relative values of the misfit are more relevant than the absolute values.

An error model that predicts a response surface that roughly preserves the shape of the exact surface may be sufficient to drive the chain to minimum misfit regions at a lower computational cost than it would be possible with the exact model alone. Sharp misfit contrasts, as the one observed around 8ft, might impair the mobility of the chain, preventing the exploration of the entire parameter space. Note, however, that in multidimensional spaces (e.g., in the full 3D parameter space of the ICF test case) sharp contrast might be less problematic than in 1D, because the higher dimension might allow the chain to bypass the misfit peak.

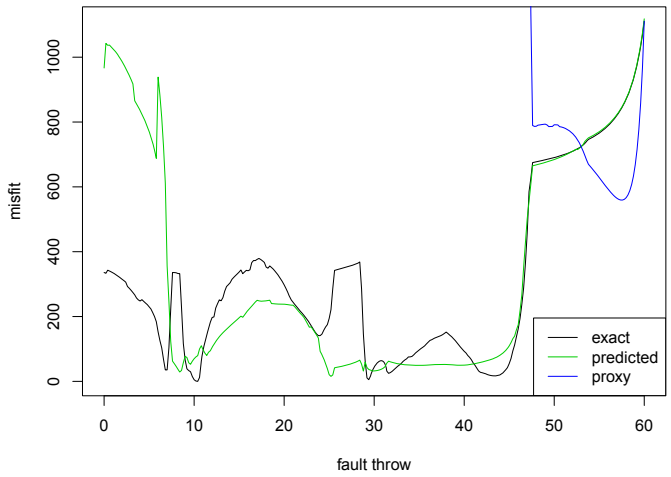


Figure 9: The 1D response surface of the ICF problem for the misfit definition given in equation 15. K_{high} and K_{low} are set to the reference values, while the fault throw varies between 0 and 60 feet. Shown are the response surfaces obtained from the exact model (black), from the responses predicted by the error model (green), and from the proxy curves alone (blue).

4.2.3. MCMC results

In a MCMC set-up the choice of proposal distribution is crucial. To obtain optimal convergence of the chain, the acceptance rate should be in the range between 20% and 50% (see Sec. 3.2). This is achieved by tuning the standard deviation of the random walk, which is defined as

$$\begin{aligned} h^{(i+1)} &= h^{(i)} + s_h \cdot {\delta_h}^{(i)}, \quad \delta_h \sim \mathcal{N}(0, \sigma^2) \\ K_h^{(i+1)} &= K_h^{(i)} + s_{K_h} \cdot {\delta_{K_h}}^{(i)}, \quad \delta_{K_h} \sim \mathcal{N}(0, \sigma^2) \\ K_l^{(i+1)} &= K_l^{(i)} + s_{K_l} \cdot {\delta_{K_l}}^{(i)}, \quad \delta_{K_l} \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{16}$$

where σ is the standard deviation of the random walk; and s_h, s_{K_h}, and s_{K_l} are the scaling factors ensuring that each prior is visited at the same rate. To determine the standard deviation that corresponds to the optimal acceptance rate for Metropolis-Hastings algorithm we have launched several chains of 1'000 iterations with different standard deviations, and found an optimal value $\sigma = 5 \cdot 10^{-3}$.

First, we compare three MH chains with three two-stage MCMC chains. All chains are launched with the optimal value $\sigma = 5 \cdot 10^{-3}$ and have a length of 10'000 iterations. The statistics of the chains are reported in table 1. A representative example of chain is plotted for each of the two methods in figure 10 (first and fourth columns). The acceptance rate of MH is approximately in the optimal interval, ranging from 14% to 36%, whereas for two-stage MCMC we obtain a slightly suboptimal acceptance rate, which ranges from 8% to 23%. In all cases the chains have been able to explore only a limited portion of the parameter space, despite a length of 10'000 iterations.

In order to enlarge the portion of the parameter space that is explored, we multiply the standard deviation

of the random walk by a factor 5 ($\sigma = 1 \cdot 10^{-2}$) and 10 ($\sigma = 5 \cdot 10^{-2}$), we launch again three chains for both values of σ. The length of the MH chains remains fixed to 10'000 iterations, whereas the length of the two-stage MCMC chains is chosen to approximately match the computational cost of the MH chains. (This is done assuming that the computational gain of the proxy with respect to the exact model is equal to the number of time steps per simulation, which is about 43). The statistics of MH and two-stage MCMC chains with the modified parameters are reported in table 1, and two examples of chains are shown in figure 10. The MH chains acceptance rate drops from an average 23% for $\sigma = 5 \cdot 10^{-3}$ to 11% and 1% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively.

In addition to the fact that these values are not optimal for convergence, the low acceptance rate implies that many of the full-physics simulations are run without providing any information gain, thus wasting computational resources. One of the main results of the work is that, at approximately the same computational cost, the two-stage MCMC set-up allows us to increase the acceptance rate by a factor 1.5 to 4 (the average acceptance is 16% and 4.5% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively, see table 1). Moreover, as the proxy model is much cheaper than the exact model, two-stage MCMC chains reach lenghts of about 15'000 and 30'000 iterations (which corresponds to an increase in length of a factor 1.5 to 3) and allows a larger portion of the parameter space to be sampled.

While those results are very promising, none of the two-stage MCMC chains visited the reference point. The reference point was visited only by one of the MH chains, which was randomly initialized very close. Overall, this test case remains very challenging for single chain MCMC set-up and multiple chains solutions (Mohamed et al., 2012) should be considered.

random walk	number of iterations			number of accepted simulations 1^{st} stage			2^{nd} stage			acceptance rate			
σ	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	mean
	Metropolis-Hasting												
$5 \cdot 10^{-3}$	10'000	10'000	10'000				1'631	3'247	1'291	18.1%	36.1%	14.3%	22.8%
$1 \cdot 10^{-2}$	10'000	10'000	10'000				1'683	755	628	18.7%	8.4%	7.0%	11.4%
$5 \cdot 10^{-2}$	10'000	10'000	10'000				179	65	48	2.0%	0.7%	0.5%	1.1%
	Two-stage MCMC												
$5 \cdot 10^{-3}$	10'000	10'000	10'000	4'760	5'299	176	367	789	41	7.7%	14.9%	23.3%	15.3%
$1 \cdot 10^{-2}$	14'372	14'815	31'738	9'666	9'656	7'820	2'060	2'075	331	23.3%	21.5%	4.2%	16.3%
$5 \cdot 10^{-2}$	28'337	31'777	27'108	9'341	9'261	9'370	393	518	337	4.2%	5.6%	3.6%	4.5%

Table 1: Results of Metropolis-Hasting and two-stage MCMC algorithms for three chains (C1, C2, and C3): the standard deviation of the random walk, σ; number of iterations (i.e. total length of the chain); the number of accepted simulations at the first-stage; the number of accepted simulations at the second-stage; and the acceptance rate (i.e., the ratio of accepted exact simulations to the number of exact simulations that have been performed).

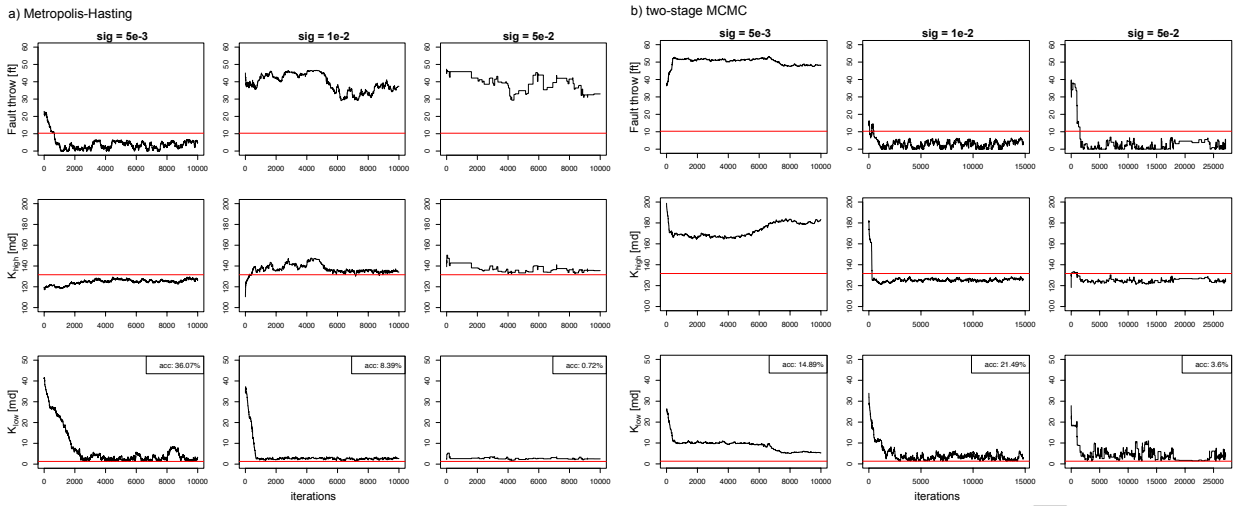


Figure 10: The chains are represented by their movements in the parameter space (vertically h, K_{high} and K_{low}) in function of iterations. For each of the three values of the random walk step length σ, one Metropolis-Hasting chain and one two-stage MCMC are plotted. The acceptance rates indicated in the legends are improved for the two-stage MCMC chains when σ is increased, allowing for much longer chains at the same computational cost.

5. Conclusions

We have investigated the potential of using error models in the context of Bayesian inference. The error model is used to map a proxy model response into the response of the exact model, which can be predicted without actually solving the exact model, thus reducing the computational costs. This methodology was applied to the ICF benchmark test case, which is geometrically simple yet very challenging. The ICF problem is particularly arduous for MCMC methods, because the very intricate surface response, characterized by sharp misfit contrasts, makes it very difficult, if not impossible, to explore the whole space by a single chain at tractable computational costs.

We have compared the performance of classic Metropolis-Hasting chains with a method that couples our error model with a two-stage MCMC algorithm. The use of the error model has increased the acceptance rate of the realizations for which the exact model was run (from 11% to 16% and 1% to 4% for $\sigma = 1 \cdot 10^{-2}$ and $\sigma = 5 \cdot 10^{-2}$, respectively). This has allowed the chain length to be increased up to a factor three with respect to MH at comparable computational costs, potentially permitting us to explore a larger portion of the response space. Based on the results of the few chains reported, it remains unclear whether the decreased computational costs might be sufficient to guide the chain out of areas of local minima, in which MCMC chains remain systematically trapped regardless of the random walk standard deviation σ that is employed. Most likely, this problem will not be solved for irregular response surfaces as the one of the ICF test case.

However, the use of an error model can be greatly beneficial also for multiple-chain algorithms that can be set up to overcome this issue.

We have demonstrated that the relationship trained on the learning set is quite effective in predicting the exact responses, as it is indicated by the correlation indices and by the linear relationships between the exact and predicted misfits. The error model has been very successful to reject bad samples, but slightly less informative to predict the response of the best samples (i.e., for realizations in regions of low misfit).

Notice that the use of the proxy without error model would be very inefficient as first-stage selection criterion. This is evident from simple inspection of the proxy misfit in figure 9: the regions of good-quality parameters cannot be identified on the basis of the proxy misfit alone. The error model is thus critical to guide the simulations in the correct regions of the parameter space, avoiding that the two-stage MCMC approach results in a counter-productive increase of simulations in poor quality regions, thus heavily increasing the computational effort.

The question that arises naturally is whether the quality of the proxy is relevant in presence of such an effective error model. To investigate this, we used the input parameters of the model (i.e., the permeabilities of the two facies and the fault throw) as proxy, that is, we directly constructed a regression model between the input parameters and the scores of the exact responses on a learning set. In this case, we have observed a total absence of relationship. This demonstrates that, despite its simplicity, the single-phase proxy provides important information on the connectivity that results from the combined effect of the parameters.

Several improvements can be devised within the framework proposed here. In particular, more complex (nonlinear) regression models could be considered (e.g., by using of kernels) and appropriate data transformations could be employed to avoid unphysical results after correction of the proxy responses, as proposed in Josset and Lunati (2013). In terms of computational cost, a major improvement could be achieved by taking advantage of all the simulations performed along the MCMC chains and iteratively updating the error model as soon as new samples are evaluated (Cui et al., 2011). This option, however, would require overcoming the problem that the likelihood is modified and convergence is not guaranteed. Several alternative approaches to MCMC could also be considered jointly with the error model, e.g., the Nested Sampling (Skilling, 2006; Elsheikh et al., 2014) in which resampling is performed at the prior level. In such approaches, the error model would be useful to reject sampled points and the Nested Sampling would avoid entrapments in the inherent structure of the ICF while allowing an iterative update of the regression model.

Acknowledgments

Many thanks are due to Pavel Tomin for his help with the flow solver, and to Imperial College and prof. J. Carter for providing the ICF data set. This project is supported by the Swiss National Science Foundation as a part of the ENSEMBLE project (Sinergia Grant No. CRSI22-132249/1) and partly by Uncertainty JIP at Heriot-Watt. The authors would like to thank the Herbette Foundation, who supported V. Demyanov's exchange with the University of Lausanne. Ivan Lunati is Swiss National Science Foundation (SNSF) Professor at the University of Lausanne (SNSF grant numbers PP00P2-123419/1 and PP00P2-144922/1).

References

Carnell R. *lhs: Latin hypercube samples.* R package version 0.5, 2009.

Carter J.N., P.J. Ballester, Z. Tavassoli, and P.R. King. *Our calibrated model has poor predictive value: An example from the petroleum industry.* Reliability Engineering & System Safety, 91(10):1373–1381, 2006.

Christen J.A., and C. Fox. *MCMC using an approximation.* Journal of Computational and Graphical statistics, 14(4):795–810, 2005.

Christie M., V. Demyanov, and D. Erbas. *Uncertainty quantification for porous media flows.* Journal of Computational Physics 217.1: 143-158, 2006.

Cui T., C. Fox, and M.J. O'Sullivan. *Adaptive Error Modelling in MCMC Sampling for Large Scale Inverse Problems.* Report, Univeristy of Auckland, Faculty of Engineering, 2011.

Demyanov V., A. Pozdnoukhov, M. Christie, and M. Kanevski. *Detection of optimal models in parameter space with support vector machines.* geoENV VII–Geostatistics for Environmental Applications, pages 345–358. Springer, 2010.

Dostert P., Y. Efendiev, and T.Y. Hou. *Multiscale finite element methods for stochastic porous media flow equations and application to uncertainty quantification.* Computer Methods in Applied Mechanics and Engineering, 197(43):3445–3455, 2008.

Efendiev Y., A. Datta-Gupta, V. Ginting, X. Ma, and B. Mallick. *An efficient two-stage Markov chain Monte Carlo method for dynamic data integration.* Water Resources Research, 41(12), 2005.

Efendiev Y., T. Hou, and W. Luo. *Preconditioning Markov chain Monte Carlo simulations using coarse-scale models.* SIAM Journal on Scientific Computing, 28(2):776–803, 2006.

Efendiev Y., A. Datta-Gupta, X. Ma and B. Mallick. *Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques.* Water Resources Research 45.11, 2009.

Elsheikh A.H., M.D. Jackson, and T.C. Laforce. *Bayesian reservoir history matching considering model and parameter uncertainties.* Mathematical Geosciences, 44(5):515–543, 2012.

Elsheikh A.H., I. Hoteit, and M.F. Wheeler. *Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates.* Computer Methods in Applied Mechanics and Engineering, 269:515–537, 2014.

Fox J., and H.S. Weisberg. *An R companion to applied regression.* Sage Publications, 2010.

Hastie T., R. Tibshirani, and J. Friedman. *The elements of statistical learning*, 2009.

Henderson B. *Exploring between site differences in water quality trends: a functional data analysis approach.* Environmetrics, 17(1):65–80, 2006.

Jenny P., S.H. Lee, and H. Tchelepi. *Multi-scale finite-volume method for elliptic problems in subsurface flow simulation.* J. Comp. Phys, 187(1):47–67, 2003.

Josset L., and I. Lunati. *Local and global error models to improve uncertainty quantification* Mathematical Geosciences, pp. 1–20, 2013.

Josset L., D. Ginsbourger and I. Lunati. *Functional error modeling for uncertainty quantification in hydrogeology.* Water Resour. Res., 51, 10501068, 2015.

Kaiser H.F. *The varimax criterion for analytic rotation in factor analysis.* Psychometrika, 23(3):187–200, 1958.

Kennedy M.C., and A. O'Hagan. *Bayesian calibration of computer models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.3: 425-464, 2001.

Laloy E., B. Rogiers, J.A. Vrugt, D. Mallants, and D. Jacques. *Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion.* Water Resources Research, 49(5):2664–2682, 2013.

Mohamed L., M.A. Christie, V. Demyanov, et al. *History matching and uncertainty quantification: multiobjective particle swarm optimisation approach.* SPE EUROPEC/EAGE Annual Conference and Exhibition. Society of Petroleum Engineers, 2011.

Mohamed L., B. Calderhead, M. Filippone, M. Christie, and M. Girolami. *Population MCMC methods for history matching and uncertainty quantification.* Computational Geosciences, 16(2):423–436, 2012.

OSullivan A.E., and M.A. Christie. *Solution error models: a new approach for coarse grid history matching.* Paper SPE, 2005.

OSullivan A.E., and M.A. Christie. *Error models for reducing history match bias.* Computational Geosciences 10.4: 405-405, 2006.

Ramsay J.O., G. Hooker, and S. Graves. *Functional data analysis with R and MATLAB.* Springer, 2009.

Ramsay J.O. *Functional data analysis.* Wiley Online Library, 2006.

Richman M.B. *Rotation of principal components.* Journal of climatology, 6(3):293–335, 1986.

Robert C.P., and G. Casella. *Monte Carlo statistical methods.* Citeseer volume 319, 2004.

Roberts G.O., A. Gelman and W.R. Gilks. *Weak convergence and optimal scaling of random walk Metropolis algorithms* The annals of applied probability, 1997.

Scheidt C., J. Caers, Y. Chen and L. Durlofsky. *Rapid Construction of Ensembles of High-resolution Reservoir Models Constrained to Production Data* 12th European Conference on the Mathematics of Oil Recovery, 2010.

Skilling, J. *Nested sampling for general Bayesian computation* Bayesian Analysis 1.4, 833-859, 2006.

Tavassoli Z., J.N. Carter, and P.R. King. *An analysis of history matching errors.* Computational Geosciences, 9(2-3):99–123, 2005.

Tavassoli Z., J.N. Carter, P.R. King, et al. *Errors in history matching.* SPE Journal, 9(03):352–361, 2004.

Weisberg S. *Applied linear regression.* John Wiley & Sons, 2014.

Zeng L., L. Shi, D. Zhang, and L. Wu. *A sparse grid based Bayesian method for contaminant source identification.* Advances in Water Resources, 37:1–9, 2012.